Print Selection

Select All | Clear | Cancel | Print | Print First Page

Section: Page(s): Print Copy: 1

Select?	Document ID	Section(s)	Page(s)	# Pages to print	Database
☑	20030097356	all	all	35	PGPB,USPT,USOC,EPAB,JPAB,DWPI
☑	6778981	all	all	N/A	PGPB,USPT,USOC,EPAB,JPAB,DWPI
☑	6134541	all	all	N/A	PGPB,USPT,USOC,EPAB,JPAB,DWPI
☑	6122628	all	all	N/A	PGPB,USPT,USOC,EPAB,JPAB,DWPI
☑	5499360	all	all	N/A	PGPB,USPT,USOC,EPAB,JPAB,DWPI

Building	Room	Printer
ran	3c18	gbrcptr

Freeform Search

Database: US Pre-Grant Publication Full-Text Database
US Patents Full-Text Database
US OCR Full-Text Database
EPO Abstracts Database
JPO Abstracts Database
Derwent World Patents Index
IBM Technical Disclosure Bulletins

Term: L1 and ((hyper near rectangle))

Display: 50 **Documents in Display Format:** - **Starting with Number** 1

Generate: ○ **Hit List** ◉ **Hit Count** ○ **Side by Side** ○ **Image**

Search Clear Interrupt

Search History

DATE: Thursday, January 06, 2005 Printable Copy Create Case

Set Name side by side	Query	Hit Count	Set Name result set
DB=PGPB,USPT,USOC,EPAB,JPAB,DWPI,TDBD; PLUR=YES; OP=OR			
L3	L1 and ((hyper near rectangle))	8	L3
L2	L1 and (partition$ near dimension)	2	L2
L1	((SEARCH$ OR RETRIEV$) near (multidimensional))	157	L1

END OF SEARCH HISTORY

DOCUMENT-IDENTIFIER: US 5499360 A
TITLE: Method for proximity searching with range testing and range adjustment

Detailed Description Text (70):
The method of FIGS. 5a-5c is chosen to maintain a constant number of comparisons at each stage of the processing "pipeline". The advantage of using rectangles, instead of circles, is that the intersection of a plurality of rectangles defined on an orthogonal coordinate system is still a rectangle I2 (as best seen in FIG. 5c). Furthermore, for a D-dimensional query, a set of hyper rectangles is used to apply the method to any number D of dimensions.

Detailed Description Text (73):
For the multi-dimensional case, the initial ranges are constructed by taking each i.sup.th coordinate of each vector value of the first attribute, and adding the maximum difference value to produce the i.sup.th maximum value. Similarly, the maximum difference value is subtracted from the i.sup.th coordinate of the vector value to produce the i.sup.th minimum value. Thus, for each value of the first attribute, a hyper rectangle is established having D minimum values and D maximum values. The total extent of the range in each direction is twice the maximum difference value. This means that adjusting the ranges does not increase the size of the range. A range either keeps the same size or grows smaller as each level of the query pipeline is executed. In this respect, the exemplary multi-dimensional embodiment differs from the exemplary one dimensional embodiment.

Detailed Description Text (81):
At step 233, after all of the attributes in the query are processed, all of the objects are processed and a determination is made at step 234 whether the current object has at least one range in the group of j.sup.th level ranges. One skilled in the art will understand that based on the testing thus far, the existence of one or more ranges in the group (or in the n.sup.th or final level group) does not necessarily indicate that the candidate tuple satisfies the query. The application of a hyper rectangle test is a screening procedure that eliminates candidate tuples. It is possible that a point falls inside the hyper-rectangle but still does not satisfy the proximity requirement.

Detailed Description Text (82):
For example, given a maximum difference of 1, some points lie within a square having sides of length 2, but are outside of a circle of diameter 2. Points that lie inside the square but outside of the circle still do not meet the proximity requirement. Consequently, at step 236, all candidate tuples that have not been screened out by the hyper rectangle test are evaluated using a Euclidean distance calculation. Such a computation is described by equation (5)

Detailed Description Text (83):
The hyper rectangle method of FIG. 6 is executed for levels 1 to j, to screen out objects that fail to satisfy the proximity criterion at any level, before performing the Euclidean distance computation of equation 5. If the proximity criterion is satisfied at step 236, then at step 238, the object is reported as being found by the query.

Detailed Description Text (84):
By use of the hyper rectangle method, most of the tuples that do not satisfy the query are screened out without performing the computationally intensive test of equation (5) for every possible tuple.

Other Reference Publication (3):
David Dobkin and Richard J. Lipton, "Multidimensional Searching Problems", Published in Siam

Journal on Computing, vol. 5, No. 2, Jun. 1976, pp. 181-186.

L3: Entry 7 of 8 File: USPT Sep 19, 2000

DOCUMENT-IDENTIFIER: US 6122628 A
TITLE: Multidimensional data clustering and dimension reduction for indexing and searching

Parent Case Text (2):
The present invention is related to patent application Ser. No. 08/961,729 entitled "Searching Multidimensional Indexes Using Associated Clustering and Dimension Reduction Information," by Castelli et al., filed of even date herewith, now pending. This co-pending application and the present invention are commonly assigned to the International Business Machines Corporation, Armonk, N.Y. This co-pending application is hereby incorporated by reference in its entirety into the present application

Brief Summary Text (22):
In accordance with the aforementioned needs, the present invention is directed to an improved apparatus and method for generating compact representations of multidimensional data. The present invention has features for generating searchable multidimensional indexes for databases. The present invention has other features for flexibly generating the indexes and for efficiently performing exact and similarity searches. The present invention has still other features for generating compact indexes which advantageously limit the amount of data transferred from disk to main memory during the search process.

Brief Summary Text (32):
In a preferred embodiment, the present invention is embodied as software stored on a program storage device readable by a machine tangibly embodying a program of instructions executable by the machine to perform method steps for generating compact representations of multidimensional data; efficiently performing exact and similarity searches; generating searchable multidimensional indexes for databases; and efficiently performing exact and similarity searches using the indexes.

Detailed Description Text (43):
FIG. 9 shows another example of a flow chart of an exact search process based on a searchable multidimensional index (108 or 612) generated according to the present invention. Here, the index (108 or 612) was generated with a recursive application of the clustering and dimensionality reduction logic. An exact search is the process of retrieving the record or records that exactly match the search template. As depicted, a query including specified data such as a search template (901) is used as input to the cluster search logic, in step 902, which is analogous to the cluster search logic in step 802 of FIG. 8. In step 902, clustering information (604) produced in step (601) of FIG. 6 is used to identify the cluster to which the search template (901) belongs. In step 903, (analogous to step 803 of FIG. 8) the dimensionality reduction information (607), produced in step 606 of FIG. 6, is used to project the input template onto the subspace associated with the cluster identified in step 902, and produce a projected template (904). In step 905, it is determined whether the current cluster is terminal, that is, if no further recursive clustering and singular value decomposition steps were applied to this cluster during the multidimensional index construction process. If the cluster is not terminal, in step 907 the search template (901) is replaced by the projected template (904), and the process returns to step 902. In step 906, if the cluster is terminal, the intra-cluster search logic uses the searchable index to search for the projected template. As noted, the simplest intra-cluster search mechanism is to conduct a linear scan (or linear search), if no spatial indexing structure can be utilized. On most occasions, spatial indexing structures such as R-trees can offer better efficiency (as compared to linear scan) when the dimension of the cluster is relatively small (smaller than 10 in most cases).

Detailed Description Text (59):
FIG. 14 shows an example of a complex surface (1401) in a 3-dimensional space and two

successive approximations (1402, 1403) based on a 3-dimensional quad tree, as taught in the art by Samet, H. in "Region Representation Quadtree from Boundary Codes" Comm. ACM 23,3, pp. 163-170 (March 1980). The first approximation (1402) is a minimal bounding box. The second approximation (1403) is the second step of a quad tree generation, where the bounding box has been divided into 8 hyper rectangles by splitting the bounding box at the midpoint of each dimension, and retaining only those hyper rectangles that intersect the surface.

Detailed Description Text (60):
In a preferred embodiment, the hierarchy of approximations is generated as a k-dimensional quad tree. An example of a method having features of the present invention for generating the hierarchy of approximations includes the steps of: generating the cluster boundaries, which correspond to a zeroth-order approximation to the geometry of the clusters; approximating the convex hull of each of the clusters by means of a minimum bounding box, thus generating a first-order approximation to the geometry of each cluster; partitioning the bounding box into 2.sup.k hyper rectangles, by cutting it at the half point of each dimension; retaining only those hyper rectangles that contain points, thus generating a second-order approximation to the geometry of the cluster; and repeating the last two steps for each of the retained hyper rectangles to generate successively the third-, fourth-, . . . , n-th approximation to the geometry of the cluster.

CLAIMS:

2. The method of claim 1, further comprising the steps of:

(a) generating cluster boundaries, which correspond to a zeroth-order approximation to a geometry of said clusters;

(b) approximating the geometry of each of the clusters by means of a minimum bounding box and generating a first-order approximation to the geometry of each cluster;

(c) partitioning the bounding box into 2 k hyper rectangles, wherein said partitioning is at a midpoint of each dimension;

(d) retaining only those hyper rectangles that contain data points and generating a second-order approximation to the geometry of the cluster; and

(e) repeating said steps (c) and (d) for each retained hyper rectangle to generate successively the third-, fourth-, . . . , n-th approximations to the geometry of the cluster.

10. The method of claim 1, for searching for k records most similar to specified data, the method for searching comprising the steps of:

identifying a cluster to which specified data belongs, based on the clustering information;

reducing the dimensionality of the specified data based on the dimensionality reduction information for an identified cluster;

generating dimensionality reduction information for reduced dimensionality specified data, in response to said reducing;

searching the multidimensional index, using the dimensionality reduction information, for a reduced-dimensionality version of the cluster to which the specified data belongs;

retrieving via the multidimensional index, the k most similar records in the cluster;

identifying other candidate clusters which can contain records closer to the specified data than the farthest of the k most similar records retrieved;

searching a closest other candidate cluster to the specified data, in response to said identifying step; and

repeating said identifying and searching steps for all said other candidate clusters.

29. The program storage device of claim 26, further comprising the steps of:

(a) generating cluster boundaries, which correspond to a zeroth-order approximation to a geometry of said clusters;

(b) approximating the geometry of each of the clusters by means of a minimum bounding box and generating a first-order approximation to the geometry of each cluster;

(c) partitioning the bounding box into 2 k hyper rectangles, wherein said partitioning is at a midpoint of each dimension;

(d) retaining only those hyper rectangles that contain data points and generating a second-order approximation to the geometry of the cluster; and

(e) repeating said steps (c) and (d) for each retained hyper rectangle to generate successively the third-, fourth-, . . . , n-th approximations to the geometry of the cluster.

37. The program storage device of claim 26, for searching for k records most similar to specified data, the method for searching comprising the steps of:

identifying a cluster to which specified data belongs, based on the clustering information;

reducing the dimensionality of the specified data based on the dimensionality reduction information for an identified cluster;

generating dimensionality reduction information for reduced dimensionality specified data, in response to said reducing;

searching the multidimensional index, using the dimensionality reduction information, for a reduced-dimensionality version of the cluster to which the specified data belongs;

retrieving via the multidimensional index, the k most similar records in the cluster;

identifying other candidate clusters which can contain records closer to the specified data than the farthest of the k most similar records retrieved;

searching a closest other candidate cluster to the specified data, in response to said determining step; and

repeating said identifying and searching steps for all said other candidate clusters.

54. The computer program product of claim 51, further comprising:

(a) means for causing a computer to effect generating cluster boundaries, which correspond to a zeroth-order approximation to a geometry of said clusters;

(b) means for causing a computer to effect approximating the geometry of each of the clusters by means of a minimum bounding box and generating a first-order approximation to the geometry of each cluster;

(c) means for causing a computer to effect partitioning the bounding box into 2 k hyper rectangles, wherein said partitioning is at a midpoint of each dimension;

(d) means for causing a computer to effect retaining only those hyper rectangles that contain data points and generating a second-order approximation to the geometry of the cluster; and

(e) means for causing a computer to effect repeating said steps (c) and (d) for each retained hyper rectangle to generate successively the third-, fourth-, . . . , n-th approximations to the geometry of the cluster.

62. The computer program product of claim 51, for searching for k records most similar to specified data, the method for searching comprising:

means for causing a computer to effect identifying a cluster to which specified data belongs, based on the clustering information;

means for causing a computer to effect reducing the dimensionality of the specified data based on the dimensionality reduction information for an identified cluster;

means for causing a computer to effect generating dimensionality reduction information for reduced dimensionality specified data, in response to said reducing;

means for causing a computer to effect searching the multidimensional index, using the dimensionality reduction information, for a reduced-dimensionality version of the cluster to which the specified data belongs;

means for causing a computer to effect retrieving via the multidimensional index, the k most similar records in the cluster;

means for causing a computer to effect identifying other candidate clusters which can contain records closer to the specified data than the farthest of the k most similar records retrieved;

means for causing a computer to effect searching a closest other candidate cluster to the specified data, in response to said identifying step; and

means for causing a computer to effect repeating said identifying and searching steps for all said other candidate clusters.

Previous Doc Next Doc Go to Doc#

L3: Entry 3 of 8 File: PGPB May 22, 2003

DOCUMENT-IDENTIFIER: US 20030097356 A1
TITLE: Apparatus and method for similarity searches using hyper-rectangle based multidimensional data segmentation

Abstract Paragraph:
Disclosed herein is an apparatus and method for similarity searches using hyper-rectangle based multidimensional data segmentation. The similarity search apparatus has MBR generation means, first sequence pruning means, second sequence pruning means, and subsequence finding means. The MBR generation means segments a multidimensional data sequence to be partitioned into subsequences, and represents each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database. The first sequence pruning means prunes irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space. The second sequence pruning means prunes irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space. The subsequence finding means detects subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Summary of Invention Paragraph:
[0002] The present invention relates generally to an apparatus and method for similarity searches using a hyper-rectangle based multidimensional data segmentation, and more particularly to an apparatus and method which can efficiently perform the segmentation with respect to data sets representable by multidimensional data sequences (MDS's), such as video streams, and can search for similarity using the segmentation.

Summary of Invention Paragraph:
[0006] As the use of multimedia data has spread to many application domains, the efficient retrieval of multidimensional, voluminous and complex information, which are the intrinsic characteristics of multimedia data, is becoming increasingly important. The present invention, as described later, belongs to retrieval technology areas for data represented by sequences, such as time-series data and multimedia data, in accordance with this retrieval requirement.

Summary of Invention Paragraph:
[0011] A query in a query process of the multidimensional sequence is given as a multidimensional sequence, and the query sequence is also divided into multiple subsequences. In one-dimensional sequence, each query subsequence is represented by a single point. However, in the multidimensional sequence, each subsequence cannot be represented by a single point, (because each point contained in each subsequence is multidimensional), such that this method cannot be used in the similarity search of the multidimensional sequence.

Summary of Invention Paragraph:
[0016] Meanwhile, a similarity search method for multidimensional data sequence, as proposed later in the present invention, uses a hyper-rectangle based segmentation, and technical fields related to the hyper-rectangle based segmentation are described as follows.

Summary of Invention Paragraph:
[0026] In accordance with one aspect of the present invention, the above object can be accomplished by the provision of an apparatus for hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising MBR generation means for segmenting a multidimensional data sequence to be

partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Summary of Invention Paragraph:
[0027] In accordance with another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance Door between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric condition determination means for determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, segment merging means for merging the next point of the sequence S.sub.i into the current segment if geometric condition is satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Summary of Invention Paragraph:
[0028] In accordance with still another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation

means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Summary of Invention Paragraph:
[0029] In accordance with still another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Summary of Invention Paragraph:
[0030] In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Summary of Invention Paragraph:
[0031] In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs

are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, merging the next point of the sequence S.sub.i into the current segment if the geometric condition is satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Summary of Invention Paragraph:
[0032] In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Summary of Invention Paragraph:
[0033] In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the

distance D.sub.norm; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Description of Drawings Paragraph:
[0035] FIG. 1 is a view showing a similarity search method for a multidimensional data sequence using hyper-rectangle based segmentation from a user's and administrator's points of view according to a preferred embodiment of the present invention;

Detail Description Paragraph:
[0053] FIG. 1 is a view showing a similarity search method for a multidimensional data sequence using hyper-rectangle based segmentation from user and administrator's points of view according to a preferred embodiment of the present invention. The above similarity search method is described in detail, as follows.

Detail Description Paragraph:
[0087] In order to measure the distance between MBRs, an MBR distance D.sub.mbr between two MBRs is introduced. Generally, MBR M in the n-dimensional Euclidean space is represented by two endpoints L (low point) and H (high point) of a major diagonal of a hyper-rectangle and defined as M=(L, H). Here, L and H are defined as L=(l.sub.1, l.sub.2, . . . , l.sub.n) and H=(h.sub.1, h.sub.2, . . . , h.sub.n), respectively, where l.sub.i.ltoreq.h.sub.i (l.ltoreq.i.ltoreq.n).

Detail Description Paragraph:
[0090] The MBR distance D.sub.mbr between two MBRs, that is, A=(L.sub.A, H.sub.A) and B=(L.sub.B, H.sub.B) in the n-dimensional Euclidean space is defined as the minimum distance between two hyper-rectangles, and defined as the following Equation 4. 4 D mbr (A , B) = (1 k n k 2) 1 / 2 x k = (| h A , k - l B , k | if h A , k < l B , k | l A , k - h B , k | if h B , k < l B , k 0 otherwise) [4]

Detail Description Paragraph:
[0145] Further, the points P.sub.j can be represented in the hyper-rectangular form for convenience by placing L.sup.1=H.sup.i=P.sup.i- .sub.j for all dimensions. That is, the MBR is represented as <P.sub.j, P.sub.j, 1>. Such a rectangle is also denoted by HR(P.sub.j) which has zero volume and edge. On the other hand, the volume Vol (HR) and the total edge length Edge (HR) of the hyper-rectangle HR are defined as the following Equation [21]. 21 Vol (HR) = 1 i n (HR H i - HR L 1) Edge (HR) = 2 n - 1 1 i n (HR H i - HR L 1) [21]

Detail Description Paragraph:
[0151] Accordingly, the volume of clusters per point VPP (HR) and the edge of clusters per point EPP (HR) of the hyper-rectangle HR are calculated as the following Equation [23]. 23 VPP (HR) = Vol (HR) HR k = 1 j n (HR H i - HR L 1) HR k EPP (HR) = Edge (HR) HR k = 2 n - 1 1 i n (HR H i - HR L 1) HR k [23]

Detail Description Paragraph:
[0152] Two hyper-rectangles can be merged during the sequence segmentation process. In order to perform this merging operation, a merging operator is defined as the following Definition 8.

Detail Description Paragraph:
[0154] The merging operator .sym. between two hyper-rectangles is defined by Equation [24] below.

Detail Description Paragraph:

[0155] According to the Definition 8, it can be recognized that the merging operator .sym. has a symmetric property. That is, HR.sub.1.sym.HR.sub.2=HR.sub.2.sym.HR.sub.1 is constructed. Consider a case that the point P is merged to the hyper-rectangle HR=<L, H, k>. This merging process will probably cause changes in the volume, the edge and the number of points of the hyper-rectangle. The amount of change in each is an important factor for clustering, and volume and edge increments can be expressed as the following Equation [25].

Detail Description Paragraph:
[0162] After the multidimensional sequences are generated from data sources, such as video clips, each sequence is partitioned into segments. The segmentation is the process for generating each segment by continuously merging a point of the sequence into a current hyper-rectangle if the predefined criteria are satisfied. Assume that a point P is merged into the hyper-rectangle HR=<L, H, k> in the unit space [0, 1].sup.n. The segmentation is done in such a way such that if the merging of the point P into the hyper-rectangle HR satisfies certain given conditions, the point P is merged into the HR of the current segment, otherwise, a new segment is started from the point P. In the segmentation process, a merging object is a hyper-rectangle, while a merged object is always a point.

Detail Description Paragraph:
[0170] Provided that a minimum hyper-rectangle containing all K points in the sequence S is HR.sub.s, a unit hyper-cube uCube is defined as a cube occupied by a single point in the space [0, 1].sup.n if all points in the sequence S are uniformly distributed over the minimum hyper-rectangle HR.sub.s. If the side-length of the cube is e, the volume and the edge are expressed as the following Equation [27]. 24 Vol (nCUBE) = e n = Vol (HR s) K Edge (uCUBE) = 2 n - 1 n e = 2 n - 1 n Vol (HR s) K n [27]

Detail Description Paragraph:
[0171] If all points of the sequence S are uniformly scattered into the space of HR.sub.s, it can be recognized that one point is allocated to a unit hyper-cube uCube. In other words, it is intuitively seen that each point of S forms a hyper-rectangle having a unit hyper-cube shape. However, the uniform distribution assumed here is not likely to occur in reality. For example, frames in a video shot are very similar, and if each frame is represented by one point, these points are clustered together. The uniform distribution provides a geometric condition for determining whether or not the merging of two clusters is allowed.

CLAIMS:

1. An apparatus for hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

2. The similarity search apparatus according to claim 1, wherein the distance D.sub.mbr is the minimum distance between two hyper-rectangles.

4. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance

D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric condition determination means for determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, segment merging means for merging the next point of the sequence S.sub.i into the current segment if geometric condition is satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

6. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

8. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance

D.sub.norm; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

10. A method for a hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

11. The similarity search method according to claim 10, wherein the distance D.sub.mbr is the minimum distance between two hyper-rectangles.

13. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation step includes the steps of: inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, merging the next point of the sequence S.sub.i into the current segment if the geometric condition is satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

15. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted

from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation step includes the steps of: inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

17. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm; wherein the MBR generation step includes the steps of: inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Previous Doc Next Doc Go to Doc#

L3: Entry 6 of 8 File: USPT Oct 17, 2000

DOCUMENT-IDENTIFIER: US 6134541 A
TITLE: Searching multidimensional indexes using associated clustering and dimension reduction information

Brief Summary Text (22):
In accordance with the aforementioned needs, the present invention is directed to an improved apparatus and method for efficiently performing exact and similarity searches on multidimensional data. One example of an application of the present invention is to multidimensional indexing. Multidimensional indexing is fundamental to spatial databases, which are widely applicable to: Geographic Information Systems (GIS); Online Analytical Processing (OLAP) for decision support using a large data warehouse; and products such as IBM's QBIC and IMAGEMINER for image mining of multimedia databases where high-dimensional feature vectors are extracted from image and video data.

Detailed Description Text (43):
FIG. 9 shows another example of a flow chart of an exact search process based on a searchable multidimensional index (108 or 612) generated according to the present invention. Here, the index (108 or 612) was generated with a recursive application of the clustering and dimensionality reduction logic. An exact search is the process of retrieving the record or records that exactly match the search template. As depicted, a query including specified data such as a search template (901) is used as input to the cluster search logic, in step 902, which is analogous to the cluster search logic in step 802 of FIG. 8. In step 902, clustering information (604) produced in step (601) of FIG. 6 is used to identify the cluster to which the search template (901) belongs. In step 903, (analogous to step 803 of FIG. 8) the dimensionality reduction information (607), produced in step 606 of FIG. 6, is used to project the input template onto the subspace associated with the cluster identified in step 902, and produce a projected template (904). In step 905, it is determined whether the current cluster is terminal, that is, if no further recursive clustering and singular value decomposition steps were applied to this cluster during the multidimensional index construction process. If the cluster is not terminal, in step 907 the search template (901) is replaced by the projected template (904), and the process returns to step 902. In step 906, if the cluster is terminal, the intra-cluster search logic uses the searchable index to search for the projected template. As noted, the simplest intra-cluster search mechanism is to conduct a linear scan (or linear search), if no spatial indexing structure can be utilized. On most occasions, spatial indexing structures such as R-trees can offer better efficiency (as compared to linear scan) when the dimension of the cluster is relatively small (smaller than 10 in most cases).

Detailed Description Text (58):
FIG. 14 shows an example of a complex surface (1401) in a 3-dimensional space and two successive approximations (1402, 1403) based on a 3-dimensional quad tree, as taught in the art by Samet, H. in "Region Representation Quadtree from Boundary Codes" Comm. ACM 23,3, pp. 163-170 (March 1980). The first approximation (1402) is a minimal bounding box. The second approximation (1403) is the second step of a quad tree generation, where the bounding box has been divided into 8 hyper rectangles by splitting the bounding box at the midpoint of each dimension, and retaining only those hyper rectangles that intersect the surface.

Detailed Description Text (59):
In a preferred embodiment, the hierarchy of approximations is generated as a k-dimensional quad tree. An example of a method having features of the present invention for generating the hierarchy of approximations includes the steps of: generating the cluster boundaries, which correspond to a zeroth-order approximation to the geometry of the clusters; approximating the convex hull of each of the clusters by means of a minimum bounding box, thus generating a first-order approximation to the geometry of each cluster, partitioning the bounding box into

2.sup.k hyper rectangles, by cutting it at the half point of each dimension; retaining only those hyper rectangles that contain points, thus generating a second-order approximation to the geometry of the cluster; and repeating the last two steps for each of the retained hyper rectangles to generate successively the third-, fourth-, . . . , n-th approximation to the geometry of the cluster.

Previous Doc Next Doc Go to Doc#

L3: Entry 4 of 8 File: USPT Aug 17, 2004

DOCUMENT-IDENTIFIER: US 6778981 B2
TITLE: Apparatus and method for similarity searches using hyper-rectangle based multidimensional data segmentation

Abstract Text (1):
Disclosed herein is an apparatus and method for similarity searches using hyper-rectangle based multidimensional data segmentation. The similarity search apparatus has MBR generation means, first sequence pruning means, second sequence pruning means, and subsequence finding means. The MBR generation means segments a multidimensional data sequence to be partitioned into subsequences, and represents each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database. The first sequence pruning means prunes irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space. The second sequence pruning means prunes irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space. The subsequence finding means detects subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Brief Summary Text (3):
The present invention relates generally to an apparatus and method for similarity searches using a hyper-rectangle based multidimensional data segmentation, and more particularly to an apparatus and method which can efficiently perform the segmentation with respect to data sets representable by multidimensional data sequences (MDS's), such as video streams, and can search for similarity using the segmentation.

Brief Summary Text (7):
As the use of multimedia data has spread to many application domains, the efficient retrieval of multidimensional, voluminous and complex information, which are the intrinsic characteristics of multimedia data, is becoming increasingly important. The present invention, as described later, belongs to retrieval technology areas for data represented by sequences, such as time-series data and multimedia data, in accordance with this retrieval requirement.

Brief Summary Text (12):
A query in a query process of the multidimensional sequence is given as a multidimensional sequence, and the query sequence is also divided into multiple subsequences. In one-dimensional sequence, each query subsequence is represented by a single point. However, in the multidimensional sequence, each subsequence cannot be represented by a single point, (because each point contained in each subsequence is multidimensional), such that this method cannot be used in the similarity search of the multidimensional sequence.

Brief Summary Text (17):
Meanwhile, a similarity search method for multidimensional data sequence, as proposed later in the present invention, uses a hyper-rectangle based segmentation, and technical fields related to the hyper-rectangle based segmentation are described as follows.

Brief Summary Text (28):
In accordance with one aspect of the present invention, the above object can be accomplished by the provision of an apparatus for hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising MBR generation means for segmenting a multidimensional data sequence to be

partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Brief Summary Text (29):
In accordance with another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance Door between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric condition determination means for determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, segment merging means for merging the next point of the sequence S.sub.i into the current segment if geometric condition is satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Summary Text (30):
In accordance with still another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation

means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Summary Text (31):
In accordance with still another aspect of the present invention, there is provided an apparatus for hyper-rectangle based multidimensional data similarity searches, comprising MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Summary Text (32):
In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

Brief Summary Text (33):
In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs

are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, merging the next point of the sequence S.sub.i into the current segment if the geometric condition is satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Summary Text (34):
In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Brief Summary Text (35):
In accordance with still another aspect of the present invention, there is provided a method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the

distance D.sub.norm ; wherein the MBR generation step includes the steps of inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Drawing Description Text (3):
FIG. 1 is a view showing a similarity search method for a multidimensional data sequence using hyper-rectangle based segmentation from a user's and administrator's points of view according to a preferred embodiment of the present invention;

Detailed Description Text (6):
FIG. 1 is a view showing a similarity search method for a multidimensional data sequence using hyper-rectangle based segmentation from user and administrator's points of view according to a preferred embodiment of the present invention. The above similarity search method is described in detail, as follows.

Detailed Description Text (41):
In order to measure the distance between MBRs, an MBR distance D.sub.mbr between two MBRs is introduced. Generally, MBR M in the n-dimensional Euclidean space is represented by two endpoints L (low point) and H (high point) of a major diagonal of a hyper-rectangle and defined as M=(L, H). Here, L and H are defined as L=(l.sub.1, l.sub.2, . . . , l.sub.n) and H=(h.sub.1, h.sub.2, . . . , h.sub.n), respectively, where l.sub.i.ltoreq.h.sub.i (l.ltoreq.i.ltoreq.n).

Detailed Description Text (44):
The MBR distance D.sub.mbr between two MBRs, that is, A=(L.sub.A, H.sub.A) and B=(L.sub.B, H.sub.B) in the n-dimensional Euclidean space is defined as the minimum distance between two hyper-rectangles, and defined as the following Equation 4. ##EQU4##

Detailed Description Text (99):
Further, the points P.sub.j can be represented in the hyper-rectangular form for convenience by placing L.sup.i =H.sup.i =P.sup.i.sub.j for all dimensions. That is, the MBR is represented as <P.sub.j, P.sub.j, 1>. Such a rectangle is also denoted by HR(P.sub.j) which has zero volume and edge. On the other hand, the volume Vol (HR) and the total edge length Edge (HR) of the hyper-rectangle HR are defined as the following Equation [21]. ##EQU21##

Detailed Description Text (105):
Accordingly, the volume of clusters per point VPP (HR) and the edge of clusters per point EPP (HR) of the hyper-rectangle HR are calculated as the following Equation [23]. ##EQU23##

Detailed Description Text (106):
Two hyper-rectangles can be merged during the sequence segmentation process. In order to perform this merging operation, a merging operator is defined as the following Definition 8.

Detailed Description Text (108):
The merging operator .sym. between two hyper-rectangles is defined by Equation [24] below.

Detailed Description Text (109):
According to the Definition 8, it can be recognized that the merging operator .sym. has a symmetric property. That is, HR.sub.1.sym.HR.sub.2 =HR.sub.2.sym.HR.sub.1 is constructed. Consider a case that the point P is merged to the hyper-rectangle HR=<L, H, k>. This merging process will probably cause changes in the volume, the edge and the number of points of the hyper-rectangle. The amount of change in each is an important factor for clustering, and volume and edge increments can be expressed as the following Equation [25].

Detailed Description Text (116):
After the multidimensional sequences are generated from data sources, such as video clips, each sequence is partitioned into segments. The segmentation is the process for generating each segment by continuously merging a point of the sequence into a current hyper-rectangle if the predefined criteria are satisfied. Assume that a point P is merged into the hyper-rectangle HR=<L, H, k> in the unit space [0, 1].sup.n. The segmentation is done in such a way such that if the merging of the point P into the hyper-rectangle HR satisfies certain given conditions, the point P is merged into the HR of the current segment, otherwise, a new segment is started from the point P. In the segmentation process, a merging object is a hyper-rectangle, while a merged object is always a point.

Detailed Description Text (124):
Provided that a minimum hyper-rectangle containing all K points in the sequence S is HR.sub.s, a unit hyper-cube uCube is defined as a cube occupied by a single point in the space [0, 1].sup.n if all points in the sequence S are uniformly distributed over the minimum hyper-rectangle HR.sub.s. If the side-length of the cube is e, the volume and the edge are expressed as the following Equation [27]. ##EQU24##

Detailed Description Text (125):
If all points of the sequence S are uniformly scattered into the space of HR.sub.s, it can be recognized that one point is allocated to a unit hyper-cube uCube. In other words, it is intuitively seen that each point of S forms a hyper-rectangle having a unit hyper-cube shape. However, the uniform distribution assumed here is not likely to occur in reality. For example, frames in a video shot are very similar, and if each frame is represented by one point, these points are clustered together. The uniform distribution provides a geometric condition for determining whether or not the merging of two clusters is allowed.

CLAIMS:

1. An apparatus for hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

2. The similarity search apparatus according to claim 1, wherein the distance D.sub.mbr is the minimum distance between two hyper-rectangles.

4. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment

minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric condition determination means for determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, segment merging means for merging the next point of the sequence S.sub.i into the current segment if geometric condition is satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

6. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

8. An apparatus for hyper-rectangle based multidimensional data similarity searches, comprising: MBR generation means for segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; first sequence pruning means for pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; second sequence pruning means for pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned by the first sequence pruning means in a multidimensional Euclidean space; and subsequence finding means for finding subsequences similar to the given query sequence by obtaining sets of points in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation means includes: threshold calculation means for inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, segment generation means for initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence

S.sub.i, geometric and semantic condition determination means for determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, segment merging means for merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and segment updating means for including the current segment in the segment set and re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

10. A method for a hyper-rectangle based multidimensional data similarity searches, the multidimensional data being representable by a multidimensional data sequence, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm.

11. The similarity search method according to claim 10, wherein the distance D.sub.mbr is the minimum distance between two hyper-rectangles.

13. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation step includes the steps of: inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume and an edge using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies a geometric condition using the bounding threshold values for the volume and the edge, merging the next point of the sequence S.sub.i into the current segment if the geometric condition is satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric condition is not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

15. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation step includes the steps of: inputting a

multidimensional sequence S.sub.i and the minimum number of points per segment minPts and calculating bounding threshold values for a volume and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

17. A method for a hyper-rectangle based multidimensional data similarity searches, comprising the steps of: segmenting a multidimensional data sequence to be partitioned into subsequences, and representing each subsequence by each Minimum Bounding Rectangle (MBR), such that sets of MBRs are generated from the multidimensional data sequence, and the MBR sets are stored in a database; pruning irrelevant data sequences using a distance D.sub.mbr between MBRs extracted from an inputted query sequence and the MBR sets stored in the database in a multidimensional Euclidean space; pruning irrelevant data sequences using a normalized distance D.sub.norm between MBRs extracted from the query sequence and the MBR sets of data sequences remaining after the data sequences are pruned in a multidimensional Euclidean space; and finding subsequences similar to the given query sequence by obtaining sets of points contained in MBRs involved in a calculation of the distance D.sub.norm from each sequence obtained using the distance D.sub.norm ; wherein the MBR generation step includes the steps of: inputting a multidimensional sequence S.sub.i and the minimum number of points per segment minPts, and calculating bounding threshold values for a volume, an edge and a semantic factor using a unit hyper-cube occupied by a single point in n-dimensional unit space, if points are uniformly distributed in a hyper-rectangle which is a minimum bounding rectangle containing all points in the sequence S.sub.i, initializing a segment set and an outlier set to empty sets and generating a current segment using a first point of the sequence S.sub.i, determining whether a next point of the sequence S.sub.i satisfies geometric and semantic conditions using the bounding threshold values for the volume, the edge and the semantic factor, merging the next point of the sequence S.sub.i into the current segment if the geometric and semantic conditions are satisfied, and including the current segment in the segment set and updating the segment set by re-generating a new current segment using the next point of the sequence S.sub.i, if the geometric and semantic conditions are not satisfied and the number of points contained in the current segment exceeds the minimum number of points per segment minPts.

Previous Doc Next Doc Go to Doc#